Heterogeneous Selection Bias

Nathan Wiseman
Department of Economics,
University of Nevada, Reno

November 16, 2015

Abstract

The Heckman [1979] selection bias model has been applied in many fields to remove bias associated with nonrandom samples. Unfortunately, the model is known to be inconsistent when there is heteroskedasticity or non-normal disturbances. While some semi-parametric estimators have been proposed, practical applications have been scarce. This is likely due to the complexity of their estimation and the inability to verify their identifying assumptions. This paper explores the implications of different sources of heteroskedasticity and more extreme heterogeneity. A recent parametric approach by Reichert and Tauchmann [2014] is considerably more simple to estimate than semi-parametric alternatives, and the Monte Carlo simulations in this paper suggest it is able to provide consistent estimates under a wide variety of data generating processes. This estimator is shown to lead to more intuitive estimates when applied to data on miles driven in order to compare the "rebound effect" of cars and trucks.

1 Introduction

The Heckman [1979] model has been used to deal with all sorts of selection bias and endogenous binary variables. It's main limitation is that it relies on the assumptions of homoskedasticity and bivariate normality between the error terms in the outcome and selection equations. The goal of this paper is to use Monte Carlo simulations to investigate the sensitivity of the Heckman [1979] model to heteroskedasticity relative to a recently proposed estimator by Reichert and Tauchmann [2014] that is meant to accommodate additional heterogeneity.

Over the years, the Heckman [1979] model has been critiqued in a variety of ways (see Puhani [2000] for a review of Monte Carlo studies):

- Sensitivity to Distributional Assumptions
- Multicollinearity of Correction Term in Two-Stage Implementations
- Minimal Improvement Over Two-Part Models
- Sensitivity to the Exclusion Restriction

Less attention has been paid to the implications of heteroskedasticity [Vella, 1998]. Most studies considering heteroskedasticity have been in the context of censored regressions, looking at the inconsistency of the Tobit Model [Hurd, 1979, Arabmazar and Schmidt, 1981, Brown and Moffitt, 1983, Brännäs and Laitila, 1989]. In the context of Tobit models for censored data, it has been suggested that violations of heteroskedasticity may lead to more bias than violations of normality [Hurd, 1979]. Fernández Sáinz et al. [1999] is the only known study to explicitly consider heteroskedasticity in sample selection models. They compared Heckman's two-step estimator to the Ahn and Powell [1993] semiparametric estimator, and found that when heteroskedasticity is present in the selection mechanism the Heckman estimator performed better, with asymptotic bias being linear in the covariance between the error terms in the selection and outcome equations.

Recent approaches to deal with heteroskedasticity in selection models include Donald [1995], Schaffner [2002], Chen and Khan [2003], Adkins and Hill [2004], and Reichert and Tauchmann [2014]. With the exception of Schaffner [2002] and Reichert and Tauchmann [2014], all previously proposed estimators have relied on semi-parametric approaches to provide consistent estimates in selection bias models with heteroskedasticity.

Donald [1995] provides a two-step method that maintains the joint normality assumption and relies on a nonparametric first stage to account for heteroskedasticity of unknown form. While the method led to improved estimates, some bias remained due to the difficulty in choosing how many terms to include in the nonparametric stage. A similar approach pioneered by Chen and Khan [2003] uses a three step procedure to address conditional heteroskedasticity in both the selection and outcome equations without making distributional assumptions. The first two steps are nonparametric estimators of the propensity score and interquartile range. The propensity score is obtained using the

Nadarya-Watson kernal density estimator, but other nonparametric techniques (e.g., Klein and Spady [1993]) could have been used. The interquartile range is estimated using quantile regression [Koenker and Bassett Jr, 1978]. The third stage uses an established relationship between the propensity score and the interquartile range to form normalized correction terms to be used in a partially linear regression. While the technique is shown to lead to consistent estimates, it still requires appropriately specifiying smoothing parameters, which can be somewhat arbitrary.

Schaffner [2002] considers two potential methods to account for heteroskedasticity in the selection equation, which stems from the fact that there are often unobserved variables that interact with observed covariates to determine the probability of selection. The first method utilizes a heteroskedastic probit for the selection equation, in which the equation for the conditional mean of the latent utility is normalized by dividing by the conditional standard deviation. A second approach recognizes that the resulting equation is highly nonlinear, and approximates the function using a polynomial with interaction terms while maintaining the homoskedasticity assumption. The results suggest that both methods lead to improved estimates, as evidenced by non-nested likelihood ratio tests [Vuong, 1989] on empirical data. The limitations of the study are that because it compares the approach on a single dataset it does not provide any theoretical justification for the efficacy or superiority of either of the approaches in general. It is possible that the improved fit, as evidenced by the LR tests, could be the result of overfitting, a common issue when using polynomial regression.

An unpublished manuscript from Adkins and Hill [2004] considers interacting polynomials similar to those considered in Schaffner [2002] with the selectivity correction term to account for heteroskedasticity. The resulting estimates are shown to have lower bias in most cases, but do not appear to be consistent estimators.

Most recently, Reichert and Tauchmann [2014] consider a selection model in which the level of the outcome variable influences the probability of selection differently for different groups. While their focus is not on heteroskedasticity, their model formulation leads to two-step and maximum likelihood estimators that control for heteroskedasticity in the selection equation resulting from the selection mechanism. It is suggested that their proposed two-step estimator can accommodate additional forms of heterogeneity, although they do not elaborate on what forms these are.

This paper will use Monte Carlo simulations to test existing parametric sample selection models under a variety of sources of heterogeneity and alternative selection mechanisms. Results suggest that the bias of the estimators increases faster when there is heteroskedasticity in the outcome equation than when it is present in the selection equation. Results also suggest that the two-step estimator suggested by Reichert and Tauchmann [2014], though sometimes inefficient, provides consistent estimates regardless of whether heteroskedasticity stems from the outcome or selection equation (or both), while the Heckman [1979] and Schaffner [2002] estimators are extremely biased when conditional

heteroskedasticity is present in the outcome equation.

2 Theoretical Selection Mechanisms

Three different model specifications were used in comparing the estimators, corresponding to the probability of selection not depending on the outcome, the probability of selection depending on the outcome homogeneously, and the probability of selection and depending on the outcome heterogeneously. The traditional selection model, in which the probability of selection does not depend on the outcome, is estimated as follows:

$$\begin{aligned} Y_i &= (\beta' X_i + \epsilon) \cdot I(Y_i^* > 0) \\ Y_i^* &= \gamma' Z_i + u_i \end{aligned}$$

Where $I(Y_i^* > 0)$ is an indicator function telling whether the latent variable, Y_i^*, is large enough such that the individual is in the selected population for which the continuous outcome, Y_i, is observed. The vector X_i contains covariates that explain the level of the outcome and the vector Z_i contains at least one additional covariate that does not directly impact the outcome (the exclusion restriction).

Traditionally Z_i also contains all of the elements of X_i, which (if the outcome equation is correctly specified) ensures there is no misspecification in the selection equation if the level of the outcome influences the probablity of selection. Consider the following data generating process:

$$\begin{aligned} Y_i &= (\beta' X_i + \epsilon_i) \cdot I(Y_i^* > 0) \\ Y_i^* &= \gamma' Z_i + \tau Y_i + u_i \end{aligned}$$

If Y_i is replaced by the RHS of the first equation,

$$\begin{aligned} Y_i^* &= \gamma' Z_i + \tau(\beta' X_i + \epsilon_i) + u_i \\ &= \gamma' Z_i + \tau\beta' X_i + (\tau\epsilon_i + u_i) \end{aligned}$$

Let $V_i = \tau\epsilon_i + u_i$ be the corresponding error term. Its variance can be calculated as:

$$\sigma_V^2 = \tau^2\sigma_\epsilon^2 + \sigma_u^2 + 2\sigma_{\epsilon u}$$

As noted by Reichert and Tauchmann [2014], due to the traditional normalization of the variance of the error term necessary for identification in probit models, the parameter τ is not identifiable. Furthermore, the traditional Heckman [1979] model will lead to consistent estimates, as the resulting error term

will also be i.i.d. normal. However, if heteroskedasticity is present in the outcome equation, the variance of the error term in the selection equation becomes:

$$\sigma^2_{(V_i|X_i)} = \tau^2\sigma^2_{(\epsilon|X_i)} + \sigma^2_u + 2\sigma_{\epsilon u}$$

Thus, the heteroskedasticity in the outcome equation carries through to the selection equation. It is also likely that with heteroskedasticity that the covariance of the error terms could vary, which would lead to the the variance of the selection equation becoming:

$$\sigma^2_{(V_i|X_i)} = \tau^2\sigma^2_{(\epsilon|X_i)} + \sigma^2_u + 2\sigma_{(\epsilon u|X_i)}$$

These circumstances are not considered by Reichert and Tauchmann [2014], but are likely to be common in cross-sectional datasets. Furthermore, Reichert and Tauchmann [2014] show that if the outcome matters differently depending on other characteristics (i.e. there are interaction effects between the level of the outcome and some demographic characteristics) then the traditional selection model becomes inconsistent due to heteroskedasticity in the selection equation. The model considered in their paper is of the form:

$$\begin{aligned} Y_i &= (\beta' X_i + \epsilon_i) \cdot I(Y_i^* > 0) \\ Y_i^* &= \gamma' Z_i + \tau Y_i + \lambda Y_i D_i + u_i \end{aligned}$$

Again replacing Y_i with the RHS of the first equation,

$$\begin{aligned} Y_i^* &= \gamma' Z_i + \tau(\beta' X_i + \epsilon_i) + \lambda(\beta' X_i + \epsilon_i)D_i + u_i \\ &= \gamma' Z_i + \tau\beta' X_i + \lambda\beta' X_i D_i + ((\tau + \lambda D_i)\epsilon_i + u_i) \end{aligned}$$

The variance of the resulting disturbance term (assuming homoskedasticity of u and ϵ) becomes

$$\sigma^2_{(V_i|X_i)} = (\tau + \lambda D_i)^2\sigma^2_\epsilon + \sigma^2_u + 2(\tau + \lambda D_i)\sigma_{\epsilon u}$$

In order to obtain consistent estimates for this model, Reichert and Tauchmann [2014] propose a two-stage estimator in which seperate selection equations are estimated for the subgroups identified by the unique combinations of D_i. The second stage utilizes the resulting propensities to construct seperate inverse mills ratios for each group, which are then multiplied by D_i and used in place of dummy variables in a linear regression for the outcome equation. This results in the following model:

$$
\begin{aligned}
Y^*_{(i|D_i=1)} &= (\gamma'_{(D_i=1)} Z_i + u_i) \cdot I(Y_i^* > 0, D_i = 1) \\
Y^*_{(i|D_i=0)} &= (\gamma'_{(D_i=0)} Z_i + u_i) \cdot I(Y_i^* > 0, D_i = 0) \\
Y_i &= \beta' X_i + E(\epsilon_i | D_i = 1) + E(\epsilon_i | D_i = 0) + \epsilon_i \\
Y_i &= \beta' X_i + \lambda(\gamma'_{(D_i=1)} Z_i) \cdot I(D_i = 1) + \lambda(\gamma'_{(D_i=1)} Z_i) \cdot I(D_i = 0) + \epsilon_i
\end{aligned}
$$

For each of the three selection mechanisms outlined above, it is possible that there could be heteroskedasticity in the outcome equation, selection equation, or both.[1] Maximum Likelihood estimators provide the most efficient estimates, but would rely on the correct specification of both the selection mechanism and any heteroskedasticity. The following section will use simulations to show the flexibility of Reichert and Tauchmann's two-step estimator relative to those of Heckman [1979] and Schaffner [2002].

3 Monte Carlo Simulations

Simulations were performed to compare OLS, the FIML and 2-step Heckman estimators, the Schaffner estimator, and the estimator proposed by Reichert & Tauchmann for the three different selection mechanisms described above. An additional hybrid estimator, combining the approaches of Schaffner and Reichert & Tauchmann, in which a heteroskedastic probit model in the first stage is combined with multiple inverse mills ratios interacted with the indicator variables in the outcome equation, was also estimated to examine whether the single first stage regression might lead to more efficient estimates. Simulations were performed with and without the various forms of heteroskedasticity in order to compare the bias and variance of the estimators. Sample size was also varied in order to get an idea of the consistency of the estimators in each case.

3.1 Simulation Methodology

The outcome equation was the same for each of the specifications:

$$Y_i = (1 + x_i + D_i + \epsilon_i) \cdot I(Y_i^* > 0)$$

However, the following three selection equations were considered:

$$
\begin{aligned}
Y_i^* &= z_i + x_i + D_i + u_i && (1) \\
Y_i^* &= z_i + x_i + D_i - Y_i + u_i && (2) \\
Y_i^* &= z_i + x_i + D_i - Y_i - 0.75 Y_i D_i + u_i && (3)
\end{aligned}
$$

[1] The model considered by Reichert and Tauchmann [2014] considers heteroskedasticity only in the selection equation, which stems from the estimation procedure necessitating replacing Y_i with $\beta' X + \epsilon$, and thus does not permit the same level of heterogeneity in the population.

Where x and z are distributed U(0,1) and D is distributed Bin(0.5). The disturbances ϵ_i and u_i follow a bivariate normal distribution with mean zero.

In all specifications, the covariance matrix, conditional on $D_i = 0$, was:

$$\begin{pmatrix} \sigma^2_{(\epsilon|D_i=0)} & \sigma_{(\epsilon u|D_i=0)} \\ \sigma_{(\epsilon u|D_i=0)} & \sigma^2_{(u|D_i=0)} \end{pmatrix} = \begin{pmatrix} 1 & .25 \\ .25 & 1 \end{pmatrix}$$

Which corresponds to $\rho_{(\epsilon u|D_i=0)} = .25$. The covariance matrix conditional on $D_i = 1$ depended on the case being considered:

Case 1: There is no heterogeneity.

Case 2: Heterogeneity in the Selection Equation

$$\begin{pmatrix} \sigma^2_{(\epsilon|D_i=1)} & \sigma_{(\epsilon u|D_i=1)} \\ \sigma_{(\epsilon u|D_i=1)} & \sigma^2_{(u|D_i=1)} \end{pmatrix} = \begin{pmatrix} 1 & .75 \\ .75 & 9 \end{pmatrix}$$

Case 3: Heterogeneity in the Outcome Equation

$$\begin{pmatrix} \sigma^2_{(\epsilon|D_i=1)} & \sigma_{(\epsilon u|D_i=1)} \\ \sigma_{(\epsilon u|D_i=1)} & \sigma^2_{(u|D_i=1)} \end{pmatrix} = \begin{pmatrix} 9 & .75 \\ .75 & 1 \end{pmatrix}$$

Case 4: Heteroskedasticity in Both the Selection and Outcome Equations

$$\begin{pmatrix} \sigma^2_{(\epsilon|D_i=1)} & \sigma_{(\epsilon u|D_i=1)} \\ \sigma_{(\epsilon u|D_i=1)} & \sigma^2_{(u|D_i=1)} \end{pmatrix} = \begin{pmatrix} 9 & 2.25 \\ 2.25 & 9 \end{pmatrix}$$

Case 5: Complete Heterogeneity

$$\begin{pmatrix} \sigma^2_{(\epsilon|D_i=1)} & \sigma_{(\epsilon u|D_i=1)} \\ \sigma_{(\epsilon u|D_i=1)} & \sigma^2_{(u|D_i=1)} \end{pmatrix} = \begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix}$$

In Cases 1-4, the correlation between the error terms is maintained ($\rho_{(\epsilon u|D_i=0)} = \rho_{(\epsilon u|D_i=1)} = .25$), while Case 5 corresponds approximately to $\rho_{(\epsilon u|D_i=1)} = -.35$.

For each of the 15 possible cases (5 cases for each of the three selection mechanisms), 1000 repetitions were performed using sample sizes of 1000 and 10000, which should help elucidate which estimators are consistent.

3.2 Simulation Results

The results of the Monte Carlo simulations are provided in Tables 1-12. Tables 1 & 2 show the bias of the estimators for 1000 and 10000 observations, respectively, when the first selection mechanism is followed (no dependence on the level of the outcome), while Tables 3 & 4 show the corresponding standard deviations. Tables 5 & 6 show the bias when the probability of selection depends homogeneously on the outcome, with standard deviations reported in Tables 7 & 8. Tables 9-12 show the results that correspond to the model considered by Reichert and Tauchmann [2014].

The results suggest that none of the estimators are superior all the time; however, the overall bias fluctuates considerably more for some of the models depending on the source of heteroskedasticity and the selection equation considered. Holding the correlation between the unobservables in the outcome and selection equations constant, heteroskedasticity in the outcome equation appears to lead to much larger bias, both in OLS and the Heckman estimators.

Table 1: Bias of Estimated Parameters ($Y_i^* = z_i + x_i + D_i + u_i$, 1000 obs)

	No Het	Sel Het	Out Het	Both Het	All het
ols x	-.0866	-.071	-.1773	-.1195	.0796
ols D	-.0916	-.0201	.0553	.2631	-.4164
heck ML x	.0029	.0012	.3313	.1341	-.0109
heck ML D	-.0016	.0101	.5053	.3387	-.4894
heck 2S x	.0036	.0045	-.029	.0105	.0339
heck 2S D	-.0009	.0118	.2054	.3172	-.4528
hetP x	.0032	-.0011	-.0293	-.0334	.0739
hetP D	-.0011	-.0033	.2059	.2834	-.4213
R T x	.003	.0004	-.0049	.0111	-.0045
R T D	0	-.0102	.0164	.0004	.0092
hybrid x	.0035	.0011	-.0021	.0116	-.0052
hybrid D	.0003	-.018	.0174	-.0067	.0099

Table 2: Bias of Estimated Parameters ($Y_i^* = z_i + x_i + D_i + u_i$ 10000 obs)

	No Het	Sel Het	Out Het	Both Het	All het
ols x	-.0876	-.0732	-.1755	-.1269	.0791
ols D	-.09	-.0182	.0508	.2648	-.4199
heck ML x	-.0008	.0024	.4434	.1589	-.0421
heck ML D	-.0022	.013	.6094	.3686	-.5169
heck 2S x	-.0007	.003	-.0223	-.0017	.0398
heck 2S D	-.0021	.0134	.2058	.3173	-.4516
hetP x	-.0007	-.0004	-.0223	-.0474	.0777
hetP D	-.0021	-.0011	.2058	.2839	-.4211
R T x	-.001	0	.0057	-.0005	.0021
R T D	-.0006	-.0037	.0054	-.0149	-.0006
hybrid x	-.0009	.0001	.0059	-.0001	.0023
hybrid D	-.0006	-.0043	.0055	-.0163	-.0006

For the first selection mechanism, the bias for the Reichert and Tauchmann [2014] 2-stage estimator is quite small regardless of the source of heterogeneity. However, it is also apparent that it is quite inefficient because the sampling distributions have very large standard deviations in some cases. In contrast

to the Heckman [1979] maximum likelihood and two-step estimates, it appears to be consistent, as the bias and standard errors become smaller with the increased sample size. The estimator proposed by Schaffner [2002] appears to lead to slightly less bias and more efficient estimates than the other estimators when there is heteroskedasticity in the selection equation, which is not surprising considering this is the context it is meant for. However, it is clear that the inconsistency of the first-stage probit model leads to much less bias than comparable heteroskedasticity in the outcome equation. The hybrid model appears to be a viable alternative to the Reichert and Tauchmann [2014] method for this selection mechanism, as it also leads to approximately unbiased estimates, regardless of the source of heterogeneity; however, the multiple first stage probit regressions appear to be preferable because the hybrid model is not always more efficient, and the average bias of the hybrid model is always slightly larger. As would be expected, Heckman's maximum likelihood estimator performs poorly when its assumptions are violated. Confirming results of Fernández Sáinz et al. [1999], Heckman's two-step estimator does well when there is heteroskedasticity in the selection equation, despite the fact that the first stage probit model is known to be inconsistent when heteroskedasticity is present. It is also interesting to note that Heckman's two-step procedure performs well for those regressors not causing heteroskedasticity.

For the second selection mechanism, the Reichert and Tauchmann [2014] model again performs well in the face of heteroskedasticity of all forms, as well as more extreme heterogeneity. In this specification its inefficiency is even more apparent. In contrast to the specification without the outcome mattering for selection, the Heckman models perform poorly when there is heteroskedasticity in the selection equation. Furthermore, when there is heteroskedasticity in the selection equation the model proposed by Schaffner no longer leads to unbiased estimates, while the hybrid model does. In fact, the hybrid model has less bias than the Reichert & Tauchmann model for some of the cases for this selection mechanism for the smaller sample size, whereas it was always slightly more biased in the previous cases. It also appears to perform at least as efficiently as the Reichert & Tauchmann model for variables not contributing to the heteroskedasticity. However, these improvements, in terms of both bias and efficiency are relatively small. For variables contributing to heteroskedasticity, the hybrid model is much less efficient than the Reichert & Tauchmann model, particularly when there is heteroskedasticity in the selection equation. This could be because with this selection mechanism the error term in the selection equation is a linear combination of homoskedastic and heteroskedastic error terms, which leads to additive rather than multiplicative heteroskedasticity.

For the final selection mechanism, OLS appears to be more biased than previous selection mechanisms when there is no heteroskedasticity. Ironically, OLS seems to perform very well in this context when there is heteroskedasticity in the selection mechanism.

Table 3: SD of Estimated Parameters ($Y_i^* = z_i + x_i + D_i + u_i$, 1000 obs)

	No Het	Sel Het	Out Het	Both Het	All het
ols x	.0739	.0735	.1807	.1775	.1207
ols D	.0788	.0856	.1654	.1788	.1203
heck ML x	.0893	.0984	.4489	.3971	.2942
heck ML D	.0958	.0905	.377	.2181	.2375
heck 2S x	.0907	.1015	.2291	.2485	.1588
heck 2S D	.0975	.0919	.2019	.1946	.1443
hetP x	.0906	.0901	.2285	.2255	.1553
hetP D	.0977	.0882	.2026	.182	.1366
R T x	.0918	.103	.2318	.2403	.1618
R T D	.1423	.3017	.2717	.77	.2341
hybrid x	.0919	.1005	.2315	.2473	.1612
hybrid D	.1426	.3065	.2711	.8003	.2347

Table 4: SD of Estimated Parameters ($Y_i^* = z_i + x_i + D_i + u_i$, 10000 obs)

	No Het	Sel Het	Out Het	Both Het	All het
ols x	.0225	.0234	.0558	.0547	.0376
ols D	.0255	.0255	.0534	.0563	.0397
heck ML x	.0281	.0317	.0906	.1546	.1028
heck ML D	.0313	.0273	.0775	.0717	.0852
heck 2S x	.0284	.0322	.0727	.0752	.0495
heck 2S D	.0316	.0274	.0674	.0596	.0466
hetP x	.0284	.0288	.0728	.0693	.0489
hetP D	.0316	.026	.0675	.0569	.0445
R T x	.0287	.0322	.0739	.0756	.0505
R T D	.0445	.085	.0848	.244	.073
hybrid x	.0288	.0322	.0739	.0749	.0499
hybrid D	.0445	.0852	.0847	.2443	.073

Table 5: Bias of Estimated Parameters ($Y_i^* = z_i + x_i + D_i - Y_i + u_i$, 1000 obs)

	No Het	Sel Het	Out Het	Both Het	All het
ols x	.0489	.0213	.1137	.0592	.0803
ols D	.0477	.3152	-1.5769	-.6921	-.8728
heck ML x	-.0004	-.005	-.0572	-.0139	-.0219
heck ML D	-.0007	.3166	-1.4669	-.6546	-.8525
heck 2S x	-.0007	-.005	-.0007	.0032	.0018
heck 2S D	-.0009	.3154	-1.6039	-.6876	-.8886
hetP x	-.0004	-.0104	.0298	.0155	.0198
hetP D	-.0009	.3283	-1.5723	-.6701	-.8662
R T x	-.0004	-.0036	-.0024	.0035	.0006
R T D	.0054	.0185	.0351	-.0233	.0076
hybrid x	-.0008	-.0045	-.0022	.0022	.0008
hybrid D	.0049	-.0097	.0729	.0826	.0411

Table 6: Bias of Estimated Parameters ($Y_i^* = z_i + x_i + D_i - Y_i + u_i$, 10000 obs)

	No Het	Sel Het	Out Het	Both Het	All het
ols x	.0489	.0257	.1137	.0566	.0767
ols D	.049	.319	-1.5838	-.6966	-.8689
heck ML x	-.0002	.0001	-.0613	-.0152	-.0262
heck ML D	-.0001	.3203	-1.4701	-.673	-.843
heck 2S x	-.0002	.0003	.0005	-.0008	.0004
heck 2S D	-.0001	.3195	-1.609	-.6908	-.8802
hetP x	-.0002	-.0059	.0301	.0122	.0171
hetP D	0	.3323	-1.5765	-.6729	-.8575
R T x	-.0002	.0007	-.0011	-.0013	-.0004
R T D	-.0005	.0025	-.0026	.0202	.0057
hybrid x	-.0002	.0008	-.0023	-.0017	-.0009
hybrid D	-.0006	.003	.0013	.0249	.0086

Table 7: SD of Estimated Parameters ($Y_i^* = z_i + x_i + D_i - Y_i + u_i$, 1000 obs)

	No Het	Sel Het	Out Het	Both Het	All het
ols x	.0718	.0716	.1275	.1574	.0962
ols D	.0842	.0829	.1468	.1882	.1158
heck ML x	.0748	.0748	.1472	.1702	.1089
heck ML D	.0862	.0834	.1695	.1992	.1284
heck 2S x	.0749	.0748	.1408	.1616	.1062
heck 2S D	.0865	.0837	.1565	.1922	.125
hetP x	.0749	.0751	.1354	.1616	.1028
hetP D	.0865	.0842	.1522	.1909	.1215
R T x	.0751	.0765	.154	.1634	.1136
R T D	.2081	.4511	.8382	1.405	.5984
hybrid x	.075	.0748	.1386	.1615	.1048
hybrid D	.2084	1.0689	.865	1.6035	.6211

Table 8: SD of Estimated Parameters ($Y_i^* = z_i + x_i + D_i - Y_i + u_i$, 10000 obs)

	No Het	Sel Het	Out Het	Both Het	All het
ols x	.0229	.024	.0427	.0515	.0311
ols D	.0259	.0274	.0466	.057	.0347
heck ML x	.0241	.025	.0481	.0576	.0354
heck ML D	.0273	.0278	.0539	.0616	.0392
heck 2S x	.0242	.025	.0459	.0535	.034
heck 2S D	.0274	.0278	.0505	.0589	.0377
hetP x	.0242	.0248	.0446	.053	.033
hetP D	.0275	.0281	.049	.0584	.0367
R T x	.0242	.0249	.0508	.0543	.0362
R T D	.065	.1263	.2326	.3875	.1801
hybrid x	.0242	.0249	.0453	.0535	.0338
hybrid D	.065	.1266	.233	.389	.1806

Table 9: Bias of Estimated Parameters ($Y_i^* = z_i + x_i + D_i - Y_i - 0.75 Y_i D_i + u_i$, 1000 obs)

	No Het	Sel Het	Out Het	Both Het	All het
ols x	-.0622	.0019	-.1041	-.0832	-.0781
ols D	-.5203	.0128	-2.304	-1.9208	-1.3976
heck ML x	-.0272	.0085	-.1305	-.0975	-.0838
heck ML D	-.3388	.0987	-1.8483	-1.6929	-1.1714
heck 2S x	-.0274	.0088	-.1162	-.0964	-.0821
heck 2S D	-.3468	.1011	-2.1623	-1.7961	-1.265
hetP x	-.0548	-.0187	-.1393	-.1176	-.1106
hetP D	-.3478	.1225	-2.194	-1.8164	-1.2817
R T x	-.0059	-.0035	-.0132	-.0124	-.0105
R T D	-.0101	.0031	-.1244	.0326	.034
hybrid x	-.0589	-.0187	-.1631	-.137	-.126
hybrid D	.0558	5.2188	181.203	97.5085	48.2349

Table 10: Bias of Estimated Parameters ($Y_i^* = z_i + x_i + D_i - Y_i - 0.75 Y_i D_i + u_i$, 10000 obs)

	No Het	Sel Het	Out Het	Both Het	All het
ols x	-.0601	.0062	-.1045	-.0823	-.0766
ols D	-.5174	.0175	-2.2947	-1.9078	-1.3977
heck ML x	-.0235	.0122	-.1323	-.098	-.0825
heck ML D	-.333	.1012	-1.8199	-1.6595	-1.1561
heck 2S x	-.0248	.0124	-.1167	-.0944	-.0791
heck 2S D	-.3462	.1022	-2.1524	-1.7846	-1.2663
hetP x	-.0528	-.0149	-.1405	-.1164	-.109
hetP D	-.3454	.1257	-2.1806	-1.8024	-1.2797
R T x	-.0001	.0007	.0013	.0005	-.0008
R T D	.0026	-.0019	.0192	.0142	.0051
hybrid x	-.056	-.0136	-.165	-.1363	-.1243
hybrid D	.006	.0043	.4034	.3452	.1994

Table 11: SD of Estimated Parameters ($Y_i^* = z_i + x_i + D_i - Y_i - 0.75 Y_i D_i + u_i$, 1000 obs)

	No Het	Sel Het	Out Het	Both Het	All het
ols x	.0734	.0786	.1145	.1302	.0889
ols D	.0883	.0893	.1355	.1615	.1072
heck ML x	.0782	.0806	.1381	.1409	.1003
heck ML D	.1021	.1023	.1909	.2326	.1309
heck 2S x	.0784	.081	.1204	.1354	.0956
heck 2S D	.1058	.1036	.1609	.1822	.1251
hetP x	.0814	.0808	.1181	.1343	.0934
hetP D	.1067	.1032	.1467	.1692	.1184
R T x	.0856	.0896	.138	.1499	.1082
R T D	.3354	.5292	2.3661	3.4049	1.2979
hybrid x	.0821	.0813	.1223	.1359	.0974
hybrid D	.4525	120.9773	1159.8886	709.7693	445.0304

Table 12: SD of Estimated Parameters ($Y_i^* = z_i + x_i + D_i - Y_i - 0.75 Y_i D_i + u_i$, 10000 obs)

	No Het	Sel Het	Out Het	Both Het	All het
ols x	.0244	.0238	.0372	.0426	.0295
ols D	.0266	.0293	.0438	.0522	.0332
heck ML x	.0257	.0244	.0455	.0471	.0333
heck ML D	.0302	.0326	.0524	.0942	.0387
heck 2S x	.0256	.0244	.0397	.0448	.0313
heck 2S D	.0309	.0329	.0498	.0583	.0385
hetP x	.0263	.0244	.0389	.0443	.0304
hetP D	.0313	.0332	.0468	.0549	.0368
R T x	.0289	.0273	.0482	.0499	.0368
R T D	.1072	.161	.531	.5891	.3207
hybrid x	.0265	.0244	.0397	.0454	.0312
hybrid D	.1146	.1946	.8943	.983	.5014

4 Empirical Application

This section applies the traditional Heckman [1979] 2-step estimator with the 2-step estimator proposed by Reichert and Tauchmann [2014] to account for heterogeneity to empirical data. The dataset used is the 2009 National Household Travel Survey (NHTS).

It has been documented that there is likely selection bias present in the estimation of the elasticity of miles driven with respect to the price per mile of driving. Most empirical analysis of the 2009 NHTS has found elasticities that seem to be unreasonable considering the results obtained from aggregate data. It is possible that the reason for this is unobserved heterogeneity in the population.

4.1 Data

The data used in the analysis comes from a relatively unexplored portion of the NHTS, which is the day trip file. This file contains seperate observations for each trip made by each person in the households surveyed. The individual trips were aggregated in order to provide a measure of the total miles driven by each vehicle within the sample on the day of the survey, which spanned 14 months from March 2008 to April 2009. It is worth mentioning that this was a unique period to collect data, as it was in the midst of the great recession. Gas prices in this period rose to record highs, over $4 per gallon, before falling below $2 a gallon less than six months later. Since most analyses of the 2009 NHTS ignore this aspect of the data and opt to use the self-reported annual mileage and estimated average annual fuel price, it is likely that their results contain considerable unseen bias. By using the daily file, it is possible to capture the extreme fluctuations in gas prices, and to control for other unobserved time varying factors, such as the economic outlook, using month dummies.

Each vehicle was classified as either a car or truck depending on the model code. Model codes above 400 were considered to be trucks, which includes pickups, vans, and large SUVs. The number of vehicle owners in the dataset, after removing observations with incomplete data and removing any household containing individuals owning more than one vehicle was 109,806.

4.2 Model

The model applied is known as a switching regression model, which is essentially when there is selection bias due to self selection into two or more categories, where the outcome variable is only observed for the category selected. In this case, the subsamples were cars and trucks. It is likely that unobserved factors could influence a person to purchase either a car or truck, such as distance to work, type of employment (e.g., those employed in construction, agriculture, etc. would be more likely to buy a truck), or family obligations (e.g., taking kids to school or sports practice often). It is also likely that these unobserved factors could vary depending on the location of residence and type of household. For

example, people in rural areas may be more likely to work in agriculture, have more kids, and work further from home, while those in urban areas may have more access to public transportation and have a greater number of amenities within walking distance. People in more densely populated areas may also be more sensitive to fuel efficiency because they are more likely to sit idle in traffic.

Seperate first-stage probit models were estimated based on 8 seperate classes which were deemed likely to have heterogenous selection bias. Using the variable hburb, households were classified based on whether they live in the following classifications: Town & Country, City, Suburb, or Urban. Since households classified into the Town & Country category are nearly evenly split between rural and urban classifications, this category was divided further by this classification. Furthermore, households within Rural Town & Country, Urban Town & Country, and City were classified by whether they are of working age or contain retired persons. Originally, Suburb and Urban were split into seperate groups for retired and working people, but these were not found to significantly alter the results and led to increased variance of the estimator. Table 13 summarizes the dispersion of the dependent variable for each group.

The model applied to the data has the form:

$$\begin{aligned} \ln(VMT_{Car})_i &= (\alpha + \beta' X_i + \gamma' D_i + \epsilon_i) \cdot I(Y_i^* > 0) \\ \ln(VMT_{Truck})_i &= (\alpha + \beta' X_i + \gamma' D_i + \epsilon_i) \cdot I(Y_i^* < 0) \\ Y_i* &= \zeta' X_i + \eta' Z_i + G(\ln(VMT_{Car}), \ln(VMT_{Truck}), D_i) + u_i \end{aligned}$$

Where the vector X_i contains the natural log of the ratio of the price of fuel to the fuel efficiency of the vehicle ($lnPTD$), the natural log of household income (lnI), the age of the driver, and indicator variables for whether the person lives in a rural area and whether the household contains at least one retired person. Additional controls were the age of the vehicle and dummies for the month and day of week for which the data was recorded. Additional covariates contained in the vector Z_i include the number of cars and the number of trucks present in the household when the vehicle was purchased, and whether the driver was a male. Instead of the vehicle age, the vehicle age at time of purchase was used in the selection equation. The function $G(\ln(VMT_{Car}), \ln(VMT_{Truck}), D_i)$ could include terms of $\ln(VMT_{Car}$, $\ln(VMT_{Truck})$, and interactions between these and D_i. By estimating the seperate first stage regressions, the interactions drop out.

4.3 Regression Results

The results for the variables of interest of the outcome regression are contained in Table 14. The Heckman [1979] and Reichert and Tauchmann [2014] estimators both lead to significantly different estimates than the combined and two-part models. Standard errors for the Reichert & Tauchmann estimator were obtained by 1000 bootstrap repetitions, as indicated by †.

Table 13: Groups for First-Stage Probit Models

	σ_{lnVMT}	observations
ret suburb	1.11	9358
working suberb	1.02	18159
ret rural TC	1.11	11279
retired urb TC	1.2	8465
working rural TC	1.03	19789
working urb TC	1.12	13888
urban	1.1	9813
city	1.15	19052

Table 14: Regression Results by Estimator, Cars

	OLS (Full)	OLS (Car)	Heckman (Car)	R & T (Car)
lnPTD	-.241	-.343	-.12	-.131
	$(.012)^{***}$	$(.018)^{***}$	$(.031)^{***}$	$(.05)^{***}_{\dagger}$
lnI	.116	.126	.134	.146
	$(.005)^{***}$	$(.007)^{***}$	$(.007)^{***}$	$(.061)^{**}_{\dagger}$
age	-.006	-.006	-.006	-.01
	$(0)^{***}$	$(0)^{***}$	$(0)^{***}$	$(.004)^{**}_{\dagger}$
rural	.373	.381	.402	.365
	$(.007)^{***}$	$(.01)^{***}$	$(.01)^{***}$	$(.08)^{***}_{\dagger}$
retired	-.106	-.125	-.14	-.089
	$(.009)^{***}$	$(.011)^{***}$	$(.011)^{***}$	$(.091)_{\dagger}$

Table 15: Regression Results by Estimator, Trucks

	OLS (Full)	OLS (Truck)	Heckman (Truck)	R & T (Truck)
lnPTD	-.241	-.268	-.383	-.188
	$(.012)^{***}$	$(.022)^{***}$	$(.039)^{***}$	$(.051)^{***}_{\dagger}$
lnI	.116	.094	.09	.142
	$(.005)^{***}$	$(.008)^{***}$	$(.008)^{***}$	$(.036)^{***}_{\dagger}$
age	-.006	-.005	-.004	-.005
	$(0)^{***}$	$(0)^{***}$	$(0)^{***}$	$(.002)^{***}_{\dagger}$
rural	.373	.345	.334	.204
	$(.007)^{***}$	$(.01)^{***}$	$(.011)^{***}$	$(.053)^{***}_{\dagger}$
retired	-.106	-.069	-.059	-.085
	$(.009)^{***}$	$(.013)^{***}$	$(.014)^{***}$	$(.057)_{\dagger}$

While the Heckman [1979] and Reichert and Tauchmann [2014] estimators agree on the "rebound effect," or elasticity of miles with respect to the price per mile, for cars being around -.13. For Trucks, however, the Heckman [1979] estimator corrects the rebound effect in the opposite direction. This is likely because there is heterogenous selection bias for trucks depending on the different groups. For example, those in rural areas and suburbs may be more likely to drive trucks due to work in agriculture or other industries that are less common in urban areas.

While the two-step and Heckman [1979] estimators suggests that the income elasticity of miles is larger for cars than trucks, the Reichert and Tauchmann [2014] estimator suggests that they both larger and of the same magnitude. Furthermore, the Reichert and Tauchmann [2014] estimator suggests that the additional miles driven in trucks by households in rural areas is considerably smaller than for cars and as estimated by the other approaches. This suggests that there is considerably different selection bias for these areas, as would be expected. In addition, the effect of having retired persons in the home is also suggested to be the same, whereas the other estimators suggest that having retired persons in the household has a larger impact on drivers of cars.

5 Conclusion

This paper explores the finite sample properties of selection bias correction procedures in the face of different sources of heteroskedasticity and different selection mechanisms. Traditional approaches, such as the Heckman MLE and 2-step estimators are often used in cross sectional datasets that are likely to have considerable individual heterogeneity, which could be manifest as different selection rules, heteroskedasticity, or both. This paper suggests that heteroskedasticity in the outcome equation is of particular concern, as the bias of traditional estimator increases much faster in this case than when comparable heteroskedasticity is present in the selection equation.

The 2-step estimator of Reichert and Tauchmann [2014] is found to perform very well under all selection processes considered and in the face of heteroskedasticity, regardless of the source. Potential limitations are that the heteroskedasticity must be groupwise (i.e., it depends on discrete covariates), and that the estimator has large variance in small samples. Preliminary results suggest that the estimator has desireable asymptotic properties, with rates of convergence approaching $\frac{1}{\sqrt{n}}$. Future research could compare the efficiency of this estimator to the seldom used semi-parametric approaches that have been proposed in the literature.

Another potentially fruitful endeavor would be to extend the two-step framework of Reichert and Tauchmann [2014] to the polychotamous choice context. This is not a trivial endeavor, as most of the instances considered in this paper would likely require a multinomial probit first-stage in order to take advantage of the properties of the normal distribution. The multinomial probit, however, could potentially lead to even greater ability to capture heterogeneity, as it

allows for correlation patterns and heteroskedasticity between selection equations that are much more general than the multinomial logit first stage utilized in most empirical applications. Such an extension could build off of Glewwe [1993]'s proposed extension of the Heckman estimator.

Another potential extension would be to extend the approach of Reichert and Tauchmann [2014] to allow for heteroskedasticity in the outcome equation that can depend on continuous regressors. This could be accomplished through a maximum likelihood estimator, similar to the one proposed in their paper. It may also be possible to discretize the continous variable of interest, or even to utilize classification tools common to machine learning in order to classify groups into latent classes based on the expected variation in the unobservables.

References

LC Adkins and R Carter Hill. Bootstrap inferences in heteroscedastic sample selection models: A monte carlo investigation. In *74th meeting of the Southern Economic Association, New Orleans, LA (November 22)*, 2004.

Hyungtaik Ahn and James L Powell. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1):3–29, 1993.

Abbas Arabmazar and Peter Schmidt. Further evidence on the robustness of the tobit estimator to heteroskedasticity. *Journal of Econometrics*, 17(2): 253–258, 1981.

Kurt Brännäs and Thomas Laitila. Heteroskedasticity in the tobit model. *Statistical Papers*, 30(1):185–196, 1989.

Charles C Brown and Robert A Moffitt. The effect of ignoring heteroscedasticity on estimates of the tobit model, 1983.

Songnian Chen and Shakeeb Khan. Semiparametric estimation of a heteroskedastic sample selection model. *Econometric Theory*, 19(06):1040–1064, 2003.

Stephen G Donald. Two-step estimation of heteroskedastic sample selection models. *Journal of Econometrics*, 65(2):347–380, 1995.

Ana Isabel Fernández Sáinz, Juan Manuel Rodríguez Poo, Inmaculada Villanúa Martín, et al. Finite sample behavior of two step estimators in selection models. 1999.

Paul Glewwe. The three-choice multinomial probit with selectivity corrections. *Econometric Theory*, 9(02):316–322, 1993.

James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

Michael Hurd. Estimation in truncated samples when there is heteroscedasticity. *Journal of Econometrics*, 11(23):247 – 258, 1979. ISSN 0304-4076.

Roger W Klein and Richard H Spady. An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421, 1993.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Patrick Puhani. The heckman correction for sample selection and its critique. *Journal of economic surveys*, 14(1):53–68, 2000.

Arndt Reichert and Harald Tauchmann. When outcome heterogeneously matters for selection: a generalized selection correction estimator. *Applied Economics*, 46(7):762–768, 2014.

Julie Anderson Schaffner. Heteroskedastic sample selection and developing-country wage equations. *Review of economics and Statistics*, 84(2):269–280, 2002.

Francis Vella. Estimating models with sample selection bias: a survey. *Journal of Human Resources*, pages 127–169, 1998.

Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.